Personalized Online Document, Image and Video Recommendation via Commodity Eye-tracking

Songhua Xu[♮,♯,‡*]

[♮]: College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang, 310027
P.R. China

Hao Jiang[‡]

[♯]: Department of Computer Science
Yale University
New Haven, Connecticut, USA, 06520-8285

Francis C.M. Lau[‡]

[‡]: Department of Computer Science
The University of Hong Kong,
Pokfulam Road, Hong Kong,
P.R. China

ABSTRACT

We propose a new recommendation algorithm for online documents, images and videos, which is personalized. Our idea is to rely on the attention time of individual users captured through commodity eye-tracking as the essential clue. The prediction of user interest over a certain online item (a document, image or video) is based on the user's attention time acquired using vision-based commodity eye-tracking during his previous reading, browsing or video watching sessions over the same type of online materials. After acquiring a user's attention times over a collection of online materials, our algorithm can predict the user's probable attention time over a new online item through data mining. Based on our proposed algorithm, we have developed a new online content recommender system for documents, images and videos. The recommendation results produced by our algorithm are evaluated by comparing with those manually labeled by users as well as by commercial search engines including Google (Web) Search, Google Image Search and YouTube.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback; H.3.7 [**Digital Libraries**]: User issues; H.5.2 [**User Interfaces**]: Input devices and strategies; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design, Experimentation, Human Factors, Measurement, Performance.

*Contact him at songhua DOT xu AT gmail DOT com.

RecSys'08, October 23–25, 2008, Lausanne, Switzerland.

Keywords

Personalized recommendation and ranking; web search; user attention; commodity eye-tracking; document, image and video recommendation; implicit user feedback.

1. INTRODUCTION

Web surfing is a part of many people's everyday life, which may include reading online documents, looking at images, watching videos, etc. Some of these online contents are the results of specific searches requested by the users; searching is thus an important operation supported by the web. Up to now, the most common searching method is keyword based, and the searching as carried out by the current generation of commercial search engines is user-independent. Recently, however, there is a growing interest in user-dependent, or personalized searching, e.g., [23, 3]. Personalized search engines need to infer user search preferences which can be derived from user feedbacks. In this paper, we propose an algorithm which can return a personalized online content recommendation according to the user's previous reading, browsing and video watching behaviors. The key feature of our algorithm is its ability to track a user's attention time over online materials, which is obtained via a data mining process and based on data captured by a vision-based commodity eye-tracking device.

As an independent thread of research, eye-tracking has attracted many researchers in the fields of human-computer interaction, user modeling, computer graphics and interactive techniques. The main advantage of eye-tracking is that it is not intrusive while acquiring user feedbacks. These feedbacks are needed for deriving users' preferences in building adaptive systems. So far, applying modern eye-tracking technologies to recommender systems has been rare. In this paper, we propose to use commodity eye-tracking to develop a personalized recommender system for online content recommendation, where the contents may include documents, images, and videos. With our commodity eye-tracking approach, we can acquire a user's attention over online materials that the user has seen, and based on which predict his attention over materials that he has not yet seen.

2. MAIN IDEAS

For a target object O_i, which could be a document, an image, or a video on the Internet, we denote the user U_j's

attention on it as $AT(O_i, U_j)$, which is the time the user spends on reading, browsing or watching the object. Let O_{i_1} and O_{i_2} be two objects of the same type, i.e., both being documents or images or videos. Without loss of generality, let's assume after they are both presented to and watched by the user U_j, we have $AT(O_{i_1}, U_j) > AT(O_{i_2}, U_j)$; then it is reasonable to infer that U_j is more interested in O_{i_1} than O_{i_2}.

By the above heuristic, to recommend an optimal list of online materials most interesting to a user, our algorithm essentially only needs to predict the attention of the user on these objects. With the prediction results, we can then return an ordered list of online materials where the ordering is by the predicted user attention times. The problem is very similar to those rating problems studied in recommender systems—i.e., given a user's rating on a number of objects, how to predict his rating on other objects which he has not yet rated.

Nevertheless, there is a major difference between the rating scenario in conventional recommender systems and our situation. The traditional rating over an item is "atomic", which means that the user gives his overall preference for the item, and not for its subcomponents or different features separately. In contrast, in our scenario, user attention is compositive, e.g., in the case of a video, the user's attention is the accumulated attention of watching a series of episodes of the video; and in the case of reading an article, the attention is the sum of the attention on every paragraph or section of the article. In fact, in reality, when we human beings form our preferences, it is often a mixed decision in which we try to balance the various sides of a choice before making the decision. We are not aware of any recommender system in practice that allows a user to specify a rating over the various components or facets of a product. This might be due to the difficulty of obtaining user feedback for the subcomponents as many users are already reluctant to provide their feedback on an object as a whole. Given the non-intrusive nature of eye-tracking for acquiring user feedback, we can obtain the user's evaluation on the subcomponents of the target object. To leverage subcomponent rating, during our algorithm design, we carefully make use of the compositive structure of user attention in inferring user preference.

The remainder of the paper is organized as follows. We first survey the most related work in Sec. 3. We explain how to acquire user attention time for documents, images and videos respectively in Sec. 4. We discuss how to infer user attention via content-based estimation in Sec. 5. Given the user attention estimation, we introduce our algorithm for personalized online visual material recommendation based on the predicted user attention in Sec. 6. We present experiment results to demonstrate the effectiveness of our method in Sec. 7. We conclude the paper and point out some future work directions in Sec. 8.

3. RELATED WORK

3.1 Personalized search engines

Personalized search engines as a relatively new track of research is drawing more and more attention these days, e.g., [23, 3]. The existing personalized search engines so far rely on user feedbacks of various kinds, which can be broadly classified into two categories—explicit and implicit; both of them can be used to infer user intentions or preferences for customizing the search [26, 30, 31]. Because users generally are least interested to provide explicit feedbacks, the trend is to derive search preferences from implicit feedbacks [8, 9, 6]. Implicit feedbacks can be quite abundant, thus ensuring the reliability of the inference. The most popular implicit user feedbacks currently utilized in commercial search systems are query history and click data.

3.2 Implicit user feedbacks

3.2.1 *Query history*

Query history probably is the most widely used implicit user feedback at present. Google's personalized search service (http://www.google.com/psearch) allows users to store their search history in their Google account which will be analyzed for personalizing their future search. In general, there exist two classes of methods for providing personalized search based on query history: those based on the whole query history of a user and those based on the query history in a particular search session. For the former, usually a user profile is maintained to describe his search preference. For example, Liu et al. [21] constructed user profiles using the whole search history through an adaptive Rocchio algorithm [14]. Speretta and Gauch [27] demonstrated that using user profiles can significantly improve search engine performance. The query history in a query session is also called the *query chain* [24]. It can be used to automatically suggest or to complete/expand a query question for a particular user based on the query history so far in the same search session [12].

3.2.2 *Click data*

Click data is another type of implicit user feedback, which has been intensively utilized, e.g., [4, 15]. The basic idea is that when a user clicks on a document, the document is considered to be of more interest to the user than the unclicked ones. There are many ways to infer user preference from click behaviors. For example, a simple approach would be when a user clicks on the i-th link in a ranked list of webpages before having clicked on any of the first $i-1$ links, we can infer that the first $i-1$ documents are no more important than the i-th document. Among the sophisticated approaches, ranking SVM algorithm [11] has been applied to find the best webpage rank according to a user click dataset [16]. In [24], cross-query preference and individual-query preference are extracted to train a webpage rank model through a ranking SVM algorithm. Sun et al. [29] proposed a method based on singular value decomposition to improve the accuracy of a recommendation system through analyzing user click data.

3.2.3 *Attention time*

Attention time, also referred to as display time or reading time, is a newly recognized type of implicit user feedbacks. It is receiving increasing popularity even though its reliability in predicting user interest has yet to be confirmed. One side of the opinion is represented by arguments made by Kelly and Belkin [18, 17], claiming that there is no reliable relationship between the interestingness of a document and its display time. In their study the display time is measured as the average reading time spent by a group of users on articles of different topics coming from the Web. The other side of the opinion, e.g., Halabi et al. [10], is that for a fixed

user in a certain query session, attention time gives a strong indication of the user interest—the more time a user spends on reading a document, the more important the document is to him. We think these conclusions are not contradicting as display time is calculated differently by the two groups. In this paper, we propose using attention time of documents, images and videos to rank these online materials, which is analogous to the attention time idea for ranking documents only. Our basic assumption is that user specific and topic specific attention times do provide a credible indication of the user's interest; based on this assumption we propose a personalized online material ranking algorithm and a recommender system based on the algorithm.

In our prior work [32], we have explored using attention time for user-oriented webpage re-ranking. Compared with that work, our new algorithm is capable of making personalized recommendation on documents, images and videos while our prior work focuses exclusively on user-oriented webpage ranking. In this paper, we also employ vision-based commodity eye-tracking as a friendly user interaction means to acquire user attention, which was not explored previously. Last but not least, the user attention times studied in this paper are per words or image region for documents and images respectively rather than for a whole document or image. This finer level of representation and analysis makes our attention time prediction more accurate and reliable.

3.2.4 Other types of implicit user feedbacks

Other types of implicit user feedbacks include display time, scrolling, annotation, bookmarking and printing behaviors. People have recently started to combine multiple types of implicit feedbacks for better inference of user interests [22]. Fox et al. [5] have made a comprehensive study and proposed a decision tree based method augmented by Bayesian modeling to infer user preference from a set of mixed types of implicit user feedbacks.

4. ACQUIRING USER ATTENTION VIA COMMODITY EYE-TRACKING

4.1 Obtaining gaze samples through vision-based commodity eye-tracking

Eye-tracking is the technology to measure either the gaze, i.e., the spot a user is looking at, or the motion of the human eyes (http://en.wikipedia.org/wiki/Eye_tracking). In our work, we use eye-tracking to measure the attention time of a user over a document, image or video through identifying the part of the screen area the user is looking at and for how long. Unfortunately, commercial eye-tracking devices are very expensive. Some researchers therefore have turned to using ordinary web cameras as eyetracking devices [19, 1, 25, 20, 28, 7]. We did the same and have assembled an eye-tracking device using a simple web camera (Logitech Quickcam Notebook Pro) and an existent eye-tracking algorithm available from the Opengazer project [33]. We additionally employed some vision techniques to create our custom eye-tracking component. This design of our eye-tracking component, or something similar, is cost effective and can be widely adopted on personal computers as many PCs these days are equipped with web cameras.

4.2 Assigning gaze samples to object segments

Through our commodity eye-tracking component, we obtain a number of fixation points on the screen, which indicate the detected gaze area of the user. For our recommender algorithm to work, we need to anchor these gaze samples onto the corresponding object segments. Object segment means a basic compositive unit of an object, e.g., a word in an article or a region in an image. By our assumption, the more gaze samples an object segment receives, the more interesting the segment is to the user. We now look at how to anchor gaze samples onto the corresponding object segments for documents, images and videos respectively. We summarize the segment definitions of different object types and their gaze-to-segment assignment methods in Table 1. This table also summarizes the user attention prediction methods for different types of objects, which will be discussed in Sec. 5.

4.2.1 Assigning gaze samples to documents

For an online document, we define its object segments as individual words. We first introduce the term "snapshot of the document" to refer to the part of the document that is displayed on the screen at the given moment. For example, if a user resizes the displaying window or scrolls to a different part of the document, a new document snapshot is said to be formed. For each snapshot of the document, we assign the gaze samples to the corresponding words in the document in a "fractional" manner. We introduce a Gaussian kernel in the assignment process. Assuming at a certain moment, the detected gaze central point is at position (x, y) in the screen space, for each word w_i that is displayed in the current document snapshot, we first compute the central displaying point of the word as the center of the bounding box of the word's displaying region. We denote it as (x_i, y_i). Then the fraction of the gaze sample to assign to the word w_i is:

$$AT(w_i) = \exp(-\frac{(x_i - x)^2}{2\sigma_x^2} - \frac{(y_i - y)^2}{2\sigma_y^2}). \quad (1)$$

The free parameters σ_x and σ_y specify how "focused" a reader scans words when reading documents. In our current implementation, we initialize σ_x and σ_y to be the average width and height of a word's displaying bounding box in the document. The overall attention that a word in the document receives is the sum of all the fractional gaze samples it is assigned in the above process. Notice that when a word occurs multiple times in the document, we accumulate all the gaze samples assigned to these occurrences. Finally, the overall attention of a user over a word is the sum of the word's attention across all the documents the user has read previously. During our processing above, we remove stop words (http://en.wikipedia.org/wiki/Stop_words) since they are not providing any substantial meaning and thus should not really have attracted the user attention. Notice that for words in the documents that are not displayed, their attention is unspecified rather than being assigned zero.

4.2.2 Assigning gaze samples to images

For an image, we define its object segments as rectangular regions in the image. How to determine these regions will be discussed shortly. Similar to our handling the case of documents above, we also use a Gaussian kernel to fractionally assign gaze samples to these rectangular image regions. For a detected gaze point whose central position is (x, y) and

Object Type	Object Segment Definition	Gaze Sample to Segment Assignment Method	User Attention Prediction Method
Document	word	inhomogeneous 2D-Gaussian (1)	based on word attention (4)
Image	rectangular image region	homogeneous 2D-Gaussian (2)	based on image region attention (7)
Video	keyframe image	linear division (3)	based on keyframe attention (8)

Table 1: Comparisons between segment definition, gaze-to-segment assignment methods and user attention prediction methods for different types of objects.

a rectangular image region m_i, we find the nearest point (x', y') within the rectangular region m_i to (x, y). Then the fraction of the gaze sample the rectangular region m_i receives is:

$$AT(m_i) = \exp(-\frac{(x_i - x)^2}{2\sigma_m^2} - \frac{(y_i - y)^2}{2\sigma_m^2}). \quad (2)$$

Here we do not differentiate between the horizontal and vertical standard deviations in the Gaussian kernel because according to our observation, when browsing images, human eyes process the visual information more equally in the vertical and horizontal directions than when reading texts. σ_m is by default set as 1 cm and can be user tuned in order to maximize the accuracy of our user attention prediction algorithm which will be presented in Sec. 5.

Now we discuss how to determine the rectangular regions in the image for constructing image segments. First, the entire image is always treated as an image segment. The number of the gaze samples the whole image receives is the total number of the gaze samples detected that fall inside the image region when the image is being displayed on the screen. And then we find the position (x_h, y_h) in the image which has the highest gaze point density. Here the density of a position is defined as the number of gaze points whose horizontal and vertical distances to the point position (x_h, y_h) are no farther than $6\sigma_m$. Once such a highest density point is detected, we test whether the total gaze sample the rectangular region receives is above a certain threshold τ. If so, we will identify the rectangular region as an image segment, whose left bottom, right bottom, right upper and left upper corners are $(x_h - 3\sigma_m, y_h - 3\sigma_m)$, $(x_h - 3\sigma_m, y_h + 3\sigma_m)$, $(x_h + 3\sigma_m, y_h + 3\sigma_m)$, and $(x_h - 3\sigma_m, y_h + 3\sigma_m)$ respectively. After that, we remove both the rectangular region from the original image as well as all the gaze samples falling into the rectangular region. We then find the next highest density point in the remaining part of the image. If its density is above τ, we will identify a new rectangular region as an image segment and continue the search process. Otherwise, our process of image segment identification terminates. Notice that the above image segment identification process is only executed when the image is not too small, i.e., larger than $6\sigma_m \times 6\sigma_m$. Otherwise, we would only treat the whole image as an image segment.

4.2.3 Assigning gaze samples to videos

Finally, for a video, its segments are simply the video keyframes. After a user watches a piece of video online, we first detect all the keyframes from the video using the keyframe detection algorithm proposed in [2]. And the detected gaze points that fall into the video displaying window for the duration of the video will be assigned to the nearest keyframes. Because most of the online videos are of a low resolution, unlike the way we are dealing with images, we do not further split the video keyframes into sub image regions. More concretely, we assume there is a gaze sample detected at time t_i which falls in the video playing window. The two nearest video keyframes, $Keyframe_+$ and $Keyframe_-$, to the time moment t_i are at time moments t_{k-} and t_{k+} respectively. Then the fractions of gaze sample that the keyframes $Keyframe_+$ and $Keyframe_-$ receive, which are denoted as $AT(Keyf_+)$ and $AT(Keyf_-)$ respectively, are computed as follows:

$$AT(Keyf_+) \triangleq \frac{|t_i - t_{k-}|}{|t_{k+} - t_{k-}|}, \quad AT(Keyf_-) \triangleq \frac{|t_i - t_{k+}|}{|t_{k+} - t_{k-}|}. \quad (3)$$

5. PREDICTION OF USER ATTENTION

Our proposed recommender algorithm deals with three types of online materials, i.e., documents, images and videos, and so we study the prediction of user attention for each type in the following.

5.1 Predicting user attention for documents

Assuming a document v_i consists of n distinct words w_1, $\cdots$, w_n, we then predict the attention of a user over the document as the average of those words whose attentions by the user are known. Formally, the user U_j's attention over document v_i is predicted as:

$$AT(v_i, U_j) \triangleq \frac{\sum_{w_k \in v_i} AT(w_k, U_j)\delta(w_k, U_j)}{\sum_{w_k \in v_i} \delta(w_k, U_j)}. \quad (4)$$

Here $\delta(w_k, U_j) = 0$ if either there is no attention sample acquired for user U_j over the word w_k, or the word w_k is a stop word. Otherwise, $\delta(w_k, U_j) = 1$. The reason we derive an average attention time here is to normalize documents with different lengths so that our predicted user attention for documents would not bias longer documents.

5.2 Predicting user attention for images

Given an image v_i consisting of n image segments, denoted as $\mathbf{v}_i \triangleq \{vs_{i,1}, vs_{i,2}, \cdots, vs_{i,n}\}$, for each of the image segment $vs_{i,j}$ $(j = 1, \cdots, n)$, we find κ image segments whose attention by the user U_j is known and which share the highest content similarity with $vs_{i,j}$. In our current experiments, κ is set as $min(10, z)$, where z is the size of the current training set, i.e., the number of image segments whose user attentions by U_j are known. We assume these κ image segments are $vs_{i,j}^l$ $(l = 1, \cdots, \kappa)$. Then we use the following equation to predict U_j's attention for $vs_{i,j}$:

$$AT(vs_{i,j}, U_j) \triangleq \frac{\sum_{l=1}^{k} \left(AT(vs_{i,j}^l, U_j)\phi^\gamma(vs_{i,j}^l, vs_{i,j})\delta(vs_{i,j}^l, vs_{i,j})\right)}{\sum_{l=1}^{k} \left(\phi^\gamma(vs_{i,j}^l, vs_{i,j})\delta(vs_{i,j}^l, vs_{i,j})\right) + \epsilon}, \quad (5)$$

where $\phi(vs_{i,j}^l, vs_{i,j})$ returns the image content similarity between $vs_{i,j}^l$ and $vs_{i,j}$. Empirically, we find the image content similarity measurement based on the feature of "Auto Color Correlogram" [13] works best in our experiments. We adapt

the code in the open source content based image retrieval library (http://www.semanticmetadata.net/lire/) for the implementation of the image similarity metric. γ is a weight controlling how the values of $\phi(\cdot,\cdot)$ will contribute to the estimation of user attention, and ϵ is a small positive number to avoid the divide-by-zero error. The function of $\delta(\cdot,\cdot)$ defined below filters out the video pairs whose similarity is below a certain threshold:

$$\delta(vs_x, vs_y) \triangleq \begin{cases} 1 & \text{If } \phi^\gamma(vs_x, vs_y) > 0.01 \\ 0 & \text{Otherwise} \end{cases}. \quad (6)$$

Using the above equation, we can predict the attention of the user U_j over all the image segments, i.e., $vs_{i,j}$ $(j = 1, \cdots, n)$. Then the overall user attention for image v_i is the maximum sum of a non-overlapping set of all its image segments, i.e.:

$$AT(v_i, U_j) \triangleq \max_{\mathbf{v}'_i} \sum_{vs_{i,j} \in \mathbf{v}'_i} AT(vs_{i,j}, U_j), \quad (7)$$

where $\mathbf{v}'_i = \{vs_{i,1}, vs_{i,2}, \cdots, vs_{i,m}\}$ is a subset of $\mathbf{v}_i = \{vs_{i,1}, vs_{i,2}, \cdots, vs_{i,n}\}$ in which $\forall\, vs_{i,x}, vs_{i,y} \in \mathbf{v}'_i, x \neq y \Rightarrow vs_{i,x} \bigcap vs_{i,y} = \emptyset$.

5.3 Predicting user attention for videos

For a video v_i consisting of n keyframes, i.e., $\mathbf{v}_i \triangleq \{vs_{i,1}, vs_{i,2}, \cdots, vs_{i,m_i}\}$, we predict its user attention as the sum of the user attention over its individual keyframes, i.e.:

$$AT(v_i, U_j) \triangleq \sum_{j=1}^{n} AT(vs_{i,j}, U_j). \quad (8)$$

To predict the user attention over a keyframe, we use a very similar approach to user attention prediction for images. The only difference is that we do not consider image segment since most online videos are not of very high resolution and thus we do not try to detect image segments for the keyframe images (see Sec. 4.2.3). More concretely, for a video keyframe image $vs_{i,j}$, we find κ video keyframe images which are most similar to $vs_{i,j}$ from videos which the user has previously watched. In this nearest neighbour search process, we also use the image similarity metric based on the feature of "Auto Color Correlogram" [13]. Assuming these κ keyframe images are $vs^l_{i,j}$ $(l = 1, \cdots, \kappa)$, then we can use (5) to predict the user U_j's attention over the keyframe $vs_{i,j}$. After U_j's attentions over all the keyframe images of video v_i are predicted, we plug them into (8) and derive our prediction over U_j's likely attention over the video v_i.

6. PERSONALIZED ONLINE CONTENT RECOMMENDATION

Now we can construct a personalized online content recommendation algorithm based on the acquired and predicted user attentions for individual users. To experiment with our algorithm, we have developed a prototype web search interface which consists of a client side for acquiring the gaze samples of individual users on different materials, i.e., documents, images and videos, and a server side for producing a personalized online content recommendation based on the prediction of users' attentions on various types of materials.

6.1 Client side

On the client side, the acquisition method mentioned in Sec. 4 is employed. The client side periodically sends the captured user gaze sample records to the server side.

6.2 Server side

The server side implements a search engine using Java. When the server side application receives a search query submitted by a user, the application will forward the query to a commercial search engine and fetch the first 300 records if they have not been previously downloaded locally. In the case of documents and images, the commercial search engine we use is Google. In the case of videos, we use YouTube as the search engine. Our search engine then predicts the user attention over each such record through the methods introduced in Sec. 5, if the attention of the user over the record is unknown. In designing our algorithm, we also take advantage of the existing ranks over these materials as produced by the commercial search engine. More concretely, we use the following equation to compute a normalized user attention offset, whose range is between 0 and 1:

$$AT_{offset}(i) = \frac{2\exp\big(-\kappa_d \cdot rank(i)\big)}{1+\exp\big(-\kappa_d \cdot rank(i)\big)}, \quad (9)$$

where $rank(i)$ denotes the rank of the material i among the 300 items retrieved by the commercial search engine. We choose such a function because it tentatively converts an item rank into a list of attention records where items ranking low in a list would receive significantly less attention. The parameter κ_d controls how sharp this dropoff is, whose typical value in our experiment is set as 0.2. Once a user U_j's attention $AT(i, U_j)$, either from sampling or prediction, and the attention offset $AT_{offset}(i)$ are known for the i-th material, we can derive the overall attention of U_j over i simply as:

$$AT_{overall}(i, U_j) \triangleq \kappa_{overall} AT(i, U_j) + AT_{offset}(i). \quad (10)$$

The parameter $\kappa_{overall}$ is a user tunable value moderating how much he would prefer the user oriented rank result to preserve the rank produced by the commercial search engine. Finally, our algorithm recommends online materials by returning a list of these items according to their respective overall user attention in descending order.

We have also implemented an automatic mechanism which sets $\kappa_{overall}$ to a low value when there are relatively few samples in the user attention training set and gradually increases the value of $\kappa_{overall}$ as the number of user attention training samples increases. The shows that our algorithm is a learning based method. However, initially, when the training set is small, like all the learning based algorithms, our algorithm suffers from the cold start problem and tends to produce inferior results. Thus we need to "borrow" the commercial search engine's item rank list while there is little data to be learned from at the beginning. In our current experiments we use the Sigmoid function to automatically vary the value of $\kappa_{overall}$ with the input of the function to be the number of documents or images or videos in the training set multiplied by a constant (typically set to 0.1). A final note regarding (10) is that the term $AT_{offset}(i)$ is uniform for all the users, which is not personalized and is produced by the commercial search engines; the intermediate user attention term $AT(i, U_j)$ and the eventual user attention term

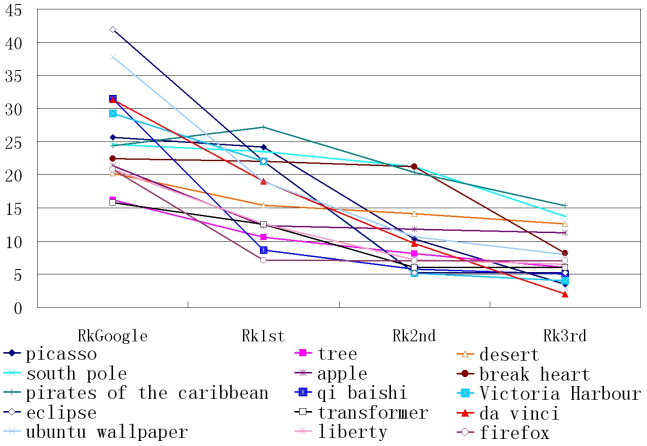

Figure 1: Plot of 15 personalized image recommendation experiment results.

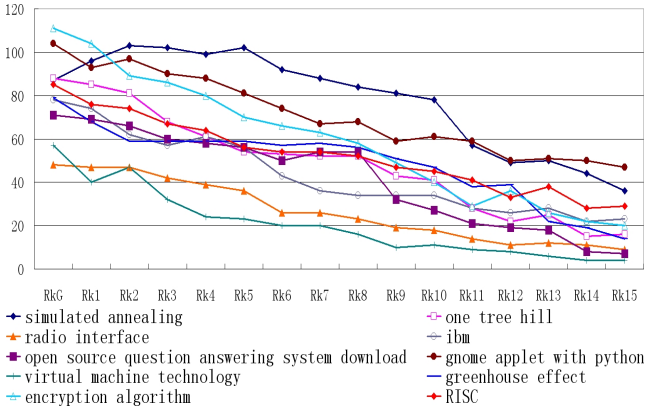

Figure 2: Plot of 10 personalized document recommendation experiment results.

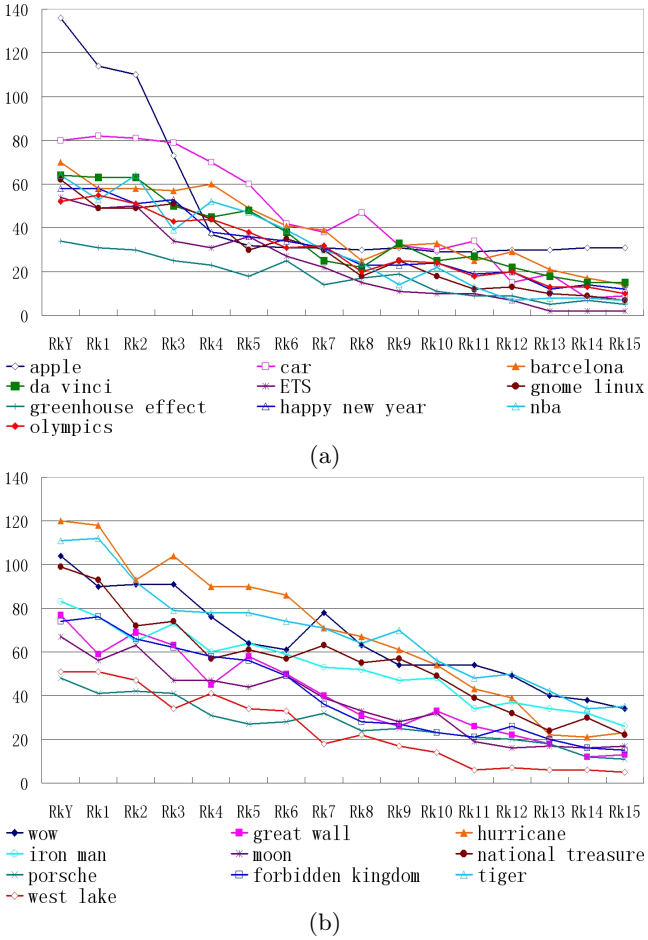

Figure 3: Plot of 20 personalized video recommendation experiment results. For easy viewing, we plot the 1st to the 10th experiment results in (a) and the 11th to the 20th experiment results in (b).

$AT_{overall}(i, U_j)$ are produced by our algorithm, which both are personalized for individual users. Because of these personalized user attention predictions, our algorithm can generate personalized online content recommendations.

7. EXPERIMENT RESULTS

We conducted our experiments for document, image and video recommendations respectively. In each experiment, the user is asked to read, browse or watch the first few documents, images or videos returned by Google or YouTube on the respective search queries. After that, he is asked to provide a rank list which reflects his interests, i.e., his expected ideal ranks over these online materials. We then use our personalized online content recommender algorithm introduced in this paper to generate a personalized recommendation list with the user attention data after he has read, browsed or watched the first i items, namely the personalized item rank list after our algorithm has access to user attention data on the first i items. We compare both Google or YouTube ranks for these items and the item ranks produced by our algorithm with respect to the user supplied groundtruth ranks.

Figures 1–3 show the results of our personalized document, image and video recommendation experiments respectively. First, in Figure 1 we show 15 results for personalized image recommendation, from an experiment involving five users. Each time a user is asked to look through the first four pages of image search results returned by Google Image Search, i.e., the top sixty image search results. After the user has browsed the first, the first two, and the first three pages of image results, our algorithm produces the personalized ranks for these images, respectively. We also ask the users to identify the images relevant to their search interest after the completion of the respective image search experiments. This information is used as groundtruth data to tell how well the Google Image Search and our algorithm perform in recommending images to Internet users. In the figure, RK_{Google}, Rk_{1st}, Rk_{2nd}, Rk_{3rd} show the average ranks of those interested images to the user in the image ranking produced by Google as well as by our algorithm after the user has browsed the first, the first two, and the first three pages of images respectively. The smaller the average rank held by these user interested images, the earlier they appear in the image search result list, which indicates a better image recommendation. In Figure 2, we show the results of our personalized document recommendation experiments. Each experiment is conducted by a different user under the same setting. Here we report the errors of each document rank

with respect to the user provided groundtruth document rank. Notice that it is the user who conducts the document search experiment that provides his most desired document rank at the end of the respective experiment. In the figure, Rk_G is the error of the initial Google rank; Rk_i is the error of the document rank produced by our algorithm after the user has read the first i documents. In all the experiments, the user is asked to read 20 documents. In computing the error, we associate the weights of {0.9, 0.9, 0.9, 0.9, 0.7, 0.7, 0.7, 0.7, 0.5, 0.5, 0.5, 0.5, 0.3, 0.3, 0.3, 0.3, 0.1, 0.1, 0.1, 0.1} with ranking errors of these 20 documents respectively so that ranking errors made with the documents appearing earlier on the recommended document list are more emphasized because they are the most important ones for a user. Figure 3 reports some experiment results for personalized video recommendation, which are conducted in a similar setting to the above document recommendation experiments except this time users are asked to watch twenty videos. In the figure, Rk_Y is the error of the initial YouTube video rank; Rk_i is the error of the video rank produced by our algorithm after the user has watched i videos. We also employ the same method to evaluate video recommendation errors with respect to the user provided groundtruth recommendation using the weighted sum of ranking errors as explained above.

In conclusion, by the results of the experiments above, we confirm that our personalized online content recommendation algorithm can indeed produce online content recommendations that are more reflective of the user's interest and preference. We expect one can save significant searching time and enjoy improved web surfing experience by adopting our proposed personalized online content recommendation algorithm.

8. CONCLUSION AND FUTURE WORK

In this paper, we propose a new personalized online content recommendation algorithm based on acquiring individual users' attention over their previously read documents, browsed images or watched videos and then predicting the users' attention over materials they have not seen through a data mining process. Due to page limit, we are only able to report some of the experiment results we have obtained. Nevertheless, the reported statistics still clearly show that our new algorithm can satisfactorily produce a personalized online content recommendation which is in better agreement with the user's expectation and preference, as verified through comparison against the benchmark algorithms by Google and YouTube. Also, having validated the domain specific prototype recommender system we have developed here using empirical results, we hope we have demonstrated the potential of employing commodity eye-tracking techniques for acquiring massive non-intrusive user feedbacks in building various types of future personalized recommender systems.

In the future, we intend to improve the precision of the image content similarity metrics by incorporating more user feedbacks. The similarity measurement is very important for producing a quality user-oriented content recommendation. In addition to exploring more existing algorithms to see whether they work well with our algorithmic framework, we also plan to study the possibility of a new similarity algorithm designed for an online learning setting. We also intend to strengthen the data mining capability of our algorithm, to optimize the performance of its implementation to better predict user preferences. Finally setting up a scalable online personalized online recommender system for massive user evaluation would be very meaningful and commercially attractive.

Acknowledgements

We thank Tao Jin for helping us in some of the experiments. The first author would like to thank David Gelernter for many inspiring discussions on intelligent web, from which the idea of this paper stems. This work has a patent pending.

9. REFERENCES

[1] T. Darrell, N. Checka, A. Oh, and L. Morency. Exploring vision-based interfaces: How to use your head in dual pointing tasks. MIT AI Memo 2002-001, 2002.

[2] F. Dirfaux. Key frame selection to represent a video. In *ICIP '00: Proceedings of IEEE International Conference on Image Processing*, volume 2, pages 275–278, 2000.

[3] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW '07: Proceedings of International Conference on World Wide Web*, pages 581–590, New York, NY, USA, 2007. ACM.

[4] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using clickthrough data and a user model. In *Query Log Analysis: Social And Technological Challenges. A Workshop at the 16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007.

[5] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2):147–168, 2005.

[6] X. Fu. Evaluating sources of implicit feedback in web searches. In *RecSys '07: Proceedings of ACM Conference on Recommender Systems*, pages 191–194, New York, NY, USA, 2007. ACM.

[7] D. Gorodnichy. Perceptual cursor–a solution to the broken loop problem in vision-based hands-free computer control devices. *National Research Council Canada Publication*, NRC-48472:1–23, 2006.

[8] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR '04: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 478–479, New York, NY, USA, 2004. ACM.

[9] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *CHI '07: Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pages 417–420, New York, NY, USA, 2007. ACM.

[10] W. S. A. Halabi, M. Kubat, and M. Tapia. Time spent on a web page is sufficient to infer a user's interest. In *IMSA '07: Proceedings of IASTED European Conference on Internet and Multimedia Systems and Applications*, pages 41–46, Anaheim, CA, USA, 2007. ACTA Press.

[11] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: an interactive retrieval evaluation and new

large test collection for research. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–201, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[12] C.-K. Huang, Y.-J. Oyang, and L.-F. Chien. A contextual term suggestion mechanism for interactive web search. In *WI '01: Proceedings of Asia-Pacific Conference on Web Intelligence: Research and Development*, pages 272–281, London, UK, 2001. Springer-Verlag.

[13] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *CVPR '97: Proceedings of International Conference on Computer Vision and Pattern Recognition*, page 762, Washington, DC, USA, 1997. IEEE Computer Society.

[14] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML '97: Proceedings of International Conference on Machine Learning*, pages 143–151, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[15] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, New York, NY, USA, 2002. ACM.

[16] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, New York, NY, USA, 2005. ACM.

[17] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 408–409, New York, NY, USA, 2001. ACM.

[18] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 377–384, New York, NY, USA, 2004. ACM.

[19] K.-N. Kim and R. Ramakrishna. Vision-based eye-gaze tracking for human computer interface. In *SMC '99: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, volume 2, pages 324–329, 1999.

[20] Y.-P. Lin, Y.-P. Chao, C.-C. Lin, and J.-H. Chen. Webcam mouse using face and eye tracking in various illumination environments. In *EMBS '05: Proceedings of 27th IEEE Annual International Conference of Engineering in Medicine and Biology Society*, pages 3738–3741, 2005.

[21] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *CIKM '02: Proceedings of International Conference on Information and Knowledge Management*, pages 558–565, New York, NY, USA, 2002. ACM.

[22] Y. Lv, L. Sun, J. Zhang, J.-Y. Nie, W. Chen, and W. Zhang. An iterative implicit feedback approach to personalized search. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 585–592, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[23] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *Communications of the ACM*, 45(9):50–55, 2002.

[24] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *KDD '05: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 239–248, New York, NY, USA, 2005. ACM.

[25] R. Ruddarraju, A. Haro, K. Nagel, Q. T. Tran, I. A. Essa, G. Abowd, and E. D. Mynatt. Perceptual user interfaces using vision-based eye tracking. In *ICMI '03: Proceedings of 5th International Conference on Multimodal Interfaces*, pages 227–233, New York, NY, USA, 2003. ACM.

[26] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Jornal of the American Society for Information Science*, 41(4):288–297, 1990.

[27] M. Speretta and S. Gauch. Personalized search based on user search histories. In *WI '05: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, pages 622–628, Washington, DC, USA, 2005. IEEE Computer Society.

[28] M.-C. Su, S.-Y. Su, and G.-D. Chen. A low-cost vision-based human-computer interface for people with severe disabilities. *Biomedical Engineering Applications, Basis, and Communications*, 17:284–292, 2005.

[29] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. Cubesvd: a novel approach to personalized web search. In *WWW '05: Proceedings of International Conference on World Wide Web*, pages 382–390, New York, NY, USA, 2005. ACM.

[30] R. White, J. M. Jose, and I. Ruthven. Comparing explicit and implicit feedback techniques for web retrieval: Trec-10 interactive track report. In *TREC*, 2001.

[31] R. White, I. Ruthven, and J. M. Jose. The use of implicit evidence for relevance feedback in web retrieval. In *Proceedings of BCS-IRSG European Colloquium on IR Research*, pages 93–109, London, UK, 2002. Springer-Verlag.

[32] S. Xu, Y. Zhu, H. Jiang, and F. C. M. Lau. A user-oriented webpage ranking algorithm based on user attention time. In *AAAI '08: Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pages 1255–1260, Chicago, USA, 2008.

[33] P. Zielinski. Opengazer: open-source gaze tracker for ordinary webcams, Samsung and The Gatsby Charitable Foundation. http://www.inference.phy.cam.ac.uk/opengazer/, 2007.